IN THE UNITED STATES PATENT AND TRADEMARK OFFICE

Patent Application

Applicant(s): L.D. Bergman et al.
Case: YOR920000742US1
Serial No.: 09/923,530
Filing Date: August 7, 2001
Group: 2172
Examiner: Baoquoc N. To

Title: Methods and Apparatus for Indexing Data in a Database and for Retrieving Data From a Database in Accordance With Queries Using Example Sets

---

DECLARATION OF PRIOR INVENTION UNDER 37 C.F.R. §1.131

We, the undersigned, hereby declare and state as follows:

1. We are the named inventors on the above-referenced U.S. patent application.

2. The invention that is the subject of the present application was conceived at least as early as October 30, 2000. On or about this date, an IBM disclosure document for an invention entitled "Indexing Method For Queries Using Multiple Positive and Negative Examples" was sent to the inventors' attorneys at the law firm of Ryan, Mason & Lewis, LLP, for preparation of a related patent application. The accompanying letter, dated October 30, 2000, from IBM in-house counsel David M. Shofi and the IBM disclosure document are attached hereto as Exhibit 1.

3. The IBM disclosure document was written by inventor Vittorio Castelli.

4. Due diligence was performed in the preparation of a patent application from the date the letter and IBM disclosure document were received until the application was filed on August 7, 2001.

5. All statements made herein of our own knowledge are true, and all statements made on information and belief are believed to be true.

6. We understand that willful false statements and the like are punishable by fine or imprisonment, or both, under 18 U.S.C. §1001, and may jeopardize the validity of the application or any patent issuing thereon.

Date: 6/24/04

Lawrence D. Bergman

Date: 6/24/04

Vittorio Castelli

Date: _____

Chung-Sheng Li

IN THE UNITED STATES PATENT AND TRADEMARK OFFICE

Patent Application

Applicant(s):	L.D. Bergman et al.
Case:	YOR920000742US1
Serial No.:	09/923,530
Filing Date:	August 7, 2001
Group:	2172
Examiner:	Baoquoc N. To

Title: Methods and Apparatus for Indexing Data in a Database and for Retrieving Data From a Database in Accordance With Queries Using Example Sets

---

DECLARATION OF PRIOR INVENTION UNDER 37 C.F.R. §1.131

We, the undersigned, hereby declare and state as follows:

1. We are the named inventors on the above-referenced U.S. patent application.

2. The invention that is the subject of the present application was conceived at least as early as October 30, 2000. On or about this date, an IBM disclosure document for an invention entitled "Indexing Method For Queries Using Multiple Positive and Negative Examples" was sent to the inventors' attorneys at the law firm of Ryan, Mason & Lewis, LLP, for preparation of a related patent application. The accompanying letter, dated October 30, 2000, from IBM in-house counsel David M. Shofi and the IBM disclosure document are attached hereto as Exhibit 1.

3. The IBM disclosure document was written by inventor Vittorio Castelli.

4. Due diligence was performed in the preparation of a patent application from the date the letter and IBM disclosure document were received until the application was filed on August 7, 2001.

5. All statements made herein of our own knowledge are true, and all statements made on information and belief are believed to be true.

6. We understand that willful false statements and the like are punishable by fine or imprisonment, or both, under 18 U.S.C. §1001, and may jeopardize the validity of the application or any patent issuing thereon.

Date: _____

_____

Lawrence D. Bergman

Date: _____

_____

Vittorio Castelli

Date: 06/22/2004

_____

Chung-Sheng Li

Thomas J. Watson Research Center
P.O. Box 218
Yorktown Heights, NY 10598

October 30, 2000

VIA OVERNIGHT MAIL

William E. Lewis, Esq.
Ryan, Mason & Lewis, LLP
90 Forrest Avenue
Locust Valley, New York 11560

-177 Subject: Preparation of Patent Application: YOR920000742US1
Yorktown Disclosure Number: YOR8-2000-0440

Title: "INDEXING METHOD FOR QUERIES USING MULTIPLE POSITIVE AND NEGATIVE EXAMPLES "

Inventor:				
	Lawrence Bergman	1-914-784-7946	1N-D20	Hawthorne
	Vittorio Castelli	1-914-945-2396	36-106D	Yorktown
	Chung-Sheng Li	1-914-784-6661	2S-D62	Hawthorne

Dear Bill,

Further to our telephone conversation of last week, enclosed are materials relative to the preparation and prosecution of the subject patent application including an original disclosure, an embodiment, 13 prior art publications and a diskette with a soft copy of the embodiment text in Lotus Word Pro format and figures in Lotus Freelance format.

The formal papers are to be prepared and filed by your office, listing the names of the Yorktown attorneys on the Declaration and Power of Attorney as follows:

POWER OF ATTORNEY: As a named inventor I hereby appoint the following attorneys and/or agents to prosecute this application and transact all business in the Patent and Trademark Office connected therewith as follows:

Manny W. Schecter (Reg. 31,722), Lauren C. Bruzzone (Reg. 35,082), Christopher A. Hughes (Reg. 26,914), Edward A. Pennington (Reg. 32,588), John E. Hoel (Reg. 26,279), Joseph C. Redmond, Jr. (Reg. 18,753), Richard M. Ludwin (Reg. 33,010), Marc A. Ehrlich (Reg. 39,966), Wayne L. Ellenbogen(Reg. 43,602), Stephen C. Kaufman (Reg 29,551), Marian Underweiser (Reg. 46,134), David M. Shofi (Reg. 39, 835), Robert M. Trepp (Reg 25,933), Louis P. Herzberg (Reg. 41,500), Louis J. Percello (Reg. 33,206), Paul J. Otterstedt (Reg. 37,411), Daniel P. Morris (Reg. 32,053), and Douglas W. Cameron (Reg 31,596).

Send correspondence to: outside counsel attorney
Direct Telephone Calls to: outside counsel attorney

After the formal papers have been prepared by your office, please send them directly to the inventors for signature with a request that they sign all the forms. As you know, notarization is no longer required by Yorktown. Please instruct the inventors to return the executed formal papers directly to your office.

Please be advised that an additional step in our procedure is required when filing all original IBM Yorktown Patent Applications in the USPTO. The additional step is that a "Taiwan Oath & Assignment" form must be completed. The form must have all the required information completely filled in and must be signed and dated by all named inventors in the subject patent application. The second page of the two page form has a section entitled, "Note 1", which contains an alphabetized code reference depicting what information must be entered into each corresponding letter field (example: (a), (b), etc.). For your convenience, enclosed on the diskette is a soft copy of the "Taiwan Oath & Assignment" form to retain for continued use when preparing all new patent applications. When the form is complete, and all inventor(s) signatures and dates have been obtained, please forward the original Taiwan Oath & Assignment back to my office.

An Associate Power of Attorney form should be prepared and forwarded to this office for signature.

Before filing, please send me, by facsimile or otherwise, a copy of the claims if not the application as a whole, for my review. Furthermore, please send me a copy of the application (hard-copy and soft-copy (WordPro97 preferable)) and the formal papers, as filed, by express mail concurrently with your filing of the application. IBM will handle any US Maintenance fee payments internally.

Please conduct the work directly with the inventors listed above, including visiting as the situation warrants. The contact inventor is Vittorio Castelli.

Please inform this office on all points involving scope of coverage and finances. Please have your illustrator prepare formal drawings for the application. In the event you file with informal drawings, please provide us with formal drawings within three months of the filing date. If you have any questions contact me at the telephone number listed below.

Best regards,

David M. Shofi
Phone: (914) 945-3247
Fax: (914) 945-3281

/jd
Enclosures

Cc:
Lauren C. Bruzzone
Barbara Rasa
Lawrence Bergman
Vittorio Castelli
Chung-Sheng Li

Note that this application involves an invention developed under govt contract.

Disclosure YOR8-2000-0440

Created By: Vittorio Castelli **Created On:** 06/02/2000 09:25:07 AM
Last Modified By: Wendy R Petrovich **Last Modified On:** 06/05/2000 08:40:55 AM

*** IBM Confidential ***

Required fields are marked with the asterisk (*) and must be filled in to complete the form .

Summary

Status *	Under Evaluation
Processing Location	YOR
Functional Area	800 Dean - Systems
Attorney/Patent Professional	David Shofi/Watson/IBM
IDT Team	David Shofi/Watson/IBM
Submitted Date	06/02/2000 10:12:14 AM
Owning Division Select	RES
	50
Incentive Program	
Lab	
Technology Code	

Inventors with Lotus Notes IDs

Inventors: Vittorio Castelli/Watson/IBM, lawrence bergman, Chung-Sheng Li/Watson/IBM@IBMUS

Inventor Name > denotes primary contact	Inventor Serial	Div/Dept	Manager Serial	Manager Name
> Castelli, Vittorio	707255	22/935E	632465	Franaszek, Peter A.
Bergman, Lawrence	660474	22/1GSA	994145	Smith, John R.
Li, Chung-Sheng	397997	22/9AUC	707138	Feldman, Stuart I.

Inventors without Lotus Notes IDs

IDT Selection

IDT Team:	Attorney/Patent Professional:
David Shofi/Watson/IBM	David Shofi/Watson/IBM

Response Due to IP&L : 07/05/2000

Main Idea

***Title of disclosure (In English)**

Indexing method for queries using multiple positive and negative examples

***Idea of disclosure**

1. Describe your invention, stating the problem solved (if appropriate), and indicating the advantages of using the invention.

In a computer system containing a database supporting similarity searches, where the user specifies similarity by providing positive

and negative examples, the search engine must translate the examples into a definition of similarity and search the repository for matches. There are at least two types of queries.

i. ``best-k-matches", where the search engine must return the k database items that more closely match the concept specified by the user.
ii. ``threshold" search, were the search engine returns all the items in the database that are more similar than a specified similarity level to the concept described by the user.

Often, the user can refine the definition of the concept by selecting good and bad examples among the returned results; this is called ``relevance feedback".
The system executes the query by creating a scoring function based on the data, and scoring the database elements.

No indexing schemes exist that support efficiently queries of type i. and ii. There are two solutions to this problem in the current art.

1. Scoring the entire database at query execution time. This solution is computationally expensive, and practically unacceptable for large databases.
2. Defining a set of ``representative" queries, and using them to modify the database in such a way that the user queries can be processed using known indexing structures. This approach is also in general not satisfactory because
   - one would have to create a new modification for each application scenario, and potentially for each user.
   - the number of "representative queries" required to modify the database appropriately could be very large.
   - the process of modifying the database using the queries is in general iterative, and hence computationally very expensive.

We propose a method for searching a static index using a branch-and-bound algorithm designed to support queries of type i. and ii specified by the user by means of positive and negative examples.

The index will be universal (i.e., it would be constructed once for a given database, and used for all used queries) and will support query-dependent scoring functions (generated from the positive and negative examples).

2. How does the invention solve the problem or achieve an advantage,(a description of "the invention", including figures inline as appropriate)?
The invention is a combination of
1. a scoring function
2. an indexing method that supports searches based on this scoring function.

Scoring Function
Our scoring function will be constructed as follow:

Let the example be a collection of M rows of the database, each having N columns. We define the column means of the example as the averages of the column values over all the samples in the example, and the column standard deviation as the standard deviation of the column values computed over all the samples in the example.

Distances are converted to scores by means of monotonically decreasing mapping functions, which have values between 0 and 1; for instance, a trapezoidal function.

When scoring an item in the database, the scores with respect to the positive and negative examples are computed, and aggregated: the maximum of the positive scores is computed, and the maximum of the negative scores is computed, and subtracted from 1. The minimum of the resulting values is then used as a score for the database object.

Indexing

Indexing algorithms supporting the above scoring function are divided into two parts: index construction and index usage. We are not concerned with index construction per se. Our invention relates to:

1. the modification of existing index methods (thus, it is very general)
2. the method for searching the modified index

The class of indexing structures to which the invention applies are based on recursive partitioning of the search space in such a way that each recursive step refines the current partition.
Such classes of methods include:

- the *R-Tree* and almost all the derived indexing structures;
- the *KD-Tree* and almost all the derived indexing structures;
- the *Quadtree* and almost all the derived indexing structures;

We describe the invention in the context of the *ordered partition*, where the database is partitioned along one dimension at a time, in a predefined number of equi-depth bins.

Modification of existing index methods

The modification to the indexing structure needed to support the method of this invention consists of dynamically attaching to each node of the tree information on the best possible score of any database item stored as a descendant of that node. For example:

1. the minimum possible distance from each positive example
2. the maximum possible distance form each negative example

Search Method

The search method, for the k-nearest-neighbor problem
consists of the following steps:

1. starting at the root, compute the score of each child of the root with respect to the whole set of examples
2. selecting the node with higher score, and breaking ties appropriately
3. recursively applying 1. and 2. until a terminal node is reached
4. searching exhaustively the terminal node, i.e., scoring all the items and inserting the k items with highest score in an ranked list
5. backtracking: visit the siblings of the current node whose score is higher than the lowest score in the ranked list, in descending order of score. Visiting means applying steps 1.,2. and 3.
6. when the last candidate sibling has been visited, repeat 5) starting from the parent of the current node, until the root is reached.

An alternative strategy consists of steps 2. to 6. with the following step substituted for step 1.

1. start at the root, for each positive example, compute the score of each child of the root with respect to that positive example and all negative examples, perform steps 2. to 6. before using the next positive example.

A further strategy consists of applying steps 1. to 3. to all the nodes of the tree except the leaves, thus computing a score for the leaves, and searching exhaustively the leaves in order of decreasing score using all positive and negative examples, until the score of next leaf is lower than the lowest score in the ordered list of results.

Some Details

We define the distance of a node with respect to an example as follows:

- The distance of a node from a positive example is the minimum possible distance of any item in the subtree rooted at the node.
- The distance of a node from a negative example is the maximum possible distance of any item in the subtree rooted at the node.

The score of a node is the score computed by applying the mapping function to the distances from all positive and negative examples.

Note that the distance of a node from a positive example is smaller than or equal to the smallest of the distances of the children of the node from the positive example. Similarly, the distance of a node from a negative example is larger than or equal to the largest of the distances of its children from the negative example. Hence, the score of a node is larger than or equal to the highestscore of its children.

Using the ordered partition, the distance between a node and an example can be quickly computed using the distance between the parent and the example, and the position of the separating point used to compute the recursive partition at the node.

3. If the same advantage or problem has been identified by others (inside/outside IBM), how have those others solved it and does your solution differ and why is it better?
The problem of searching based on multiple examples is becoming the standard in multimedia databases. We do not know of any solution that uses indexes to solve the problem.

4. If the invention is implemented in a product or prototype, include technical details, purpose, disclosure details to others and the date of that implementation.
The invention has been implemented, however the code is just a research prototype: it has not been advertised internally, nor disclosed externally.